1970-04

Analysis of a cohort prediction model with applications to student enrollment forecasting

Hager, Raymond David Jr.

Monterey, California; Naval Postgraduate School

ANALYSIS OF A COHORT PREDICTION MODEL WITH APPLICATIONS TO STUDENT ENROLLMENT FORECASTING

Raymond David Hager

United States
Naval Postgraduate School

THESIS

ANALYSIS OF A COHORT PREDICTION MODEL WITH APPLICATIONS TO STUDENT ENROLLMENT FORECASTING

by

Raymond David Hager, Jr.

April 1970

This document has been approved for public release and sale; its distribution is unlimited.

Analysis of a Cohort Prediction Model with Applications to Student Enrollment Forecasting

by

Raymond David Hager, Jr.
Lieutenant Commander, United States Navy
B.S., United States Naval Academy, 1959

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
April 1970

ABSTRACT

A model is presented for the prediction of future organization size based on the numbers of recruits entering the organization in the past. This model utilizes the correlation between populations of successive time periods in order to better estimate the future remaining personnel in the system. The number of personnel leaving from each recruit cohort is assumed to follow the same probability distribution, which is a function of the age of the cohort in the organization and the grade in which the cohort started. For large cohort sizes the total personnel in the system is approximately normally distributed. This result justifies the use of a best linear prediction method which takes into account past errors of estimating the continuing population from one period to the next. Sensitivity of predictions to errors in probability estimates is discussed. The model is applied to predicting university student enrollment. Comparison of predicted and actual student enrollment is

TABLE OF CONTENTS

I. INTRODUCTION ---------- 7

II. MODEL ---------- 11

III. MEETING SPECIFIED GOALS ---------- 22

IV. SENSITIVITY TO ERRORS IN PROBABILITY ESTIMATES ---------- 30

V. APPLICATION TO UNIVERSITY ENROLLMENT ---------- 34

APPENDIX A: REMAINDER TERMS TO PREDICTION ERROR ---------- 40

BIBLIOGRAPHY ---------- 42

INITIAL DISTRIBUTION LIST ---------- 43

FORM DD 1473 ---------- 45

LIST OF TABLES

I. New Enrollments at the University of California, Berkeley ---------- 35

II. Fractions of Students Attending Each Successive Fall After Enrollment ---------- 36

III. Values Used in Prediction ---------- 37

IV. Predicted Total Fall Enrollment ---------- 39

I. INTRODUCTION

In many large corporations and institutions with a high rate of personnel turnover, a crucial problem in recruitment planning is that of predicting from one period to the next, how many personnel presently in the organization will remain. When the length of service of a single member is fixed, or completely controlled by the organization, the problem is trivial. However, when the length of service is variable, such as with middle management in large corporations and the military service, or university student bodies, probabilistic arguments must be used to estimate expected attrition.

The theory of Markov chains has been widely used in prediction models. A basic assumption in such models is one of stationarity of rates of movement within a system of defined states. In order to assess the transition probabilities between various states in the system, it is necessary to identify various characteristics of personnel in the organization in the past to make predictions in the future. A number of such models can be found in the literature, including Bartholomew [1967], and Thonstad [1968].

Of particular interest is a paper by McAfee [1970], which describes a different type of prediction model, the so-called cohort model. For this model McAfee tests the stationarity properties of the distribution of the remaining

fraction of an initial cohort size in subsequent periods after entry into the organization. This was done for three different cohorts which entered the system in three adjacent periods. This property will be assumed to hold in the prediction model of this paper. When the cohort sizes vary from period to period, McAfee showed that the Markov Models alluded to above do not accurately describe movement of personnel in the system. Since later in this paper we consider new cohort sizes as control variables, it is not meaningful to consider them constant in size over time, and hence we concentrate in this paper on the cohort model and analyze some of its characteristics.

A basic assumption of this cohort model is that all members of a given cohort behave independently of each other, and each member's lifetime in the system is a realization from a common stochastic process. These assumptions lead to the binomial distribution for predicting the continuing fraction of a cohort with a given age in the system.

The cohort prediction model views the number of personnel in the system as a superposition of continuing portions of past cohorts. We assume that the behavior of a member in one cohort is independent of, but from the same distribution as, that of a member in a different cohort. (By the word different, is meant different time of entry into the system.) To estimate the continuing portion of the present organization size for the next period, we sum the continuing portions of the past cohorts, basing the

expected number from each cohort on the size at entry and age in the organization. (viz., the sum of expected values of independent random variables.)

As time elapses in the system for a given cohort, the number remaining in the organization from one period to the next is dependent on the number remaining from the previous period. There exists a correlation, which is positive, between the remaining portion of a single cohort in one period with the remaining portion of that same cohort in the previous period. This correlation is cumulative when we examine the correlation between the sum of continuing portions in one period with the sum of the continuing portions of the past period. It is this correlation property of the model which will be of special significance in improving the prediction characteristics of the model.

Another property of the model which will be proved and used to advantage is that with large cohort sizes, the distribution of the sum of continuing cohort portions asymptotically approaches a normal distribution. This allows us to derive simple tractable formulae for predicting the number in the system in a given time period, given the number present in the previous time period, with no detailed knowledge of from which cohort the various members came. Without this property, the exact expression for the expected value of the organization size next period, given the

realization of the present size, is intractible. A best linear predictor for the expected value of the organization size next period is this conditional expectation when the organization size is a normal random variable.

Mathematical expressions for the prediction of future organization size, knowing the size of each past and present cohort are derived. A decomposition of the organization into grades is then made to predict future grade sizes within the organization. Finally, an application of the model is made to the university student enrollment problem, where the desired prediction is that of total enrollment using data on past periods for new enrollments. This application is made with data from the University of California, Berkeley, during the period 1961 to 1969. For this model there are 16 lifetime distributions which repeat yearly; one each for freshman, sophomore, junior and senior new students admitted into each of four academic quarters.

II. MODEL

Denote by $X_i(u)$ the number of persons who enter an organization at time u in state i. Let K be the number of grades in the organization and M be the number of time periods (epochs) beyond which no member can remain in the organization. Let the probability of a single member remaining until at least s epochs have elapsed (since entry into the system), be $p_i(s)$, and let $X_i(u,s)$ be the number who entered in grade i at time u who are still in the organization at time u+s. Then $X_i(t-s,s)$ is the number in the system at time t of those that entered at time t-s in grade i, a binomially distributed random variable with parameters $X_i(t-s)$ and $p_i(s)$. The expected value of $X_i(t-s,s)$ is $p_i(s)\cdot X_i(t-s)$ and the variance is $X_i(t-s)p_i(s)q_i(s)$, where $q_i(s)$ is $1-p_i(s)$. When the time elapsed since cohort entry, s, reaches the value M, the expected value and the variance of $X_i(t-M,M)$ are both zero. It is assumed that the behavior of members of a cohort which entered the system at time t, is independent of those in a cohort which entered at time u, $u \neq t$. It is also assumed that for $i \neq j$, the behavior of the members of $X_i(t)$ is unaffected by that of members of $X_j(t)$.

Denote by $Y_i(t)$ the number of persons present at time t who started in grade i, and by Y(t) the entire population present in the organization at time t. The entire

organization size, $Y(t)$, at time t, is the sum of all remaining portions of cohort sizes which started in each of the K grades, counting back M epochs from the present time. Thus we have that

$$Y(t) = \sum_{i=1}^{K} Y_i(t) = \sum_{i=1}^{K} \sum_{s=0}^{M} X_i(t-s,s) .$$

The expected value and variance of $Y(t)$ are then respectively,

$$E(Y(t)) = \sum_{i=1}^{K} \sum_{s=0}^{M} X_i(t-s)\, p_i(s)$$

and

$$Var(Y(t)) = \sum_{i=1}^{K} \sum_{s=0}^{M} X_i(t-s)\, p_i(s)\, q_i(s) ,$$

where $p_i(0) = 1$ and $q_i(0) = 0$.

Our objective is to find expressions for the conditional expectations, $E[Y(t+1)|Y(t)]$ and $E[Y_i(t+1)|Y_i(t)]$, for which we need the distributions of $Y(t)$ and $Y_i(t)$. Although the first and second moments of $Y_i(t)$ and $Y(t)$ have simple forms, explicit formulae for their distribution functions are very unwieldy. However, if all initial cohort sizes are large for all time t, then the distributions for $Y_i(t)$ and $Y(t)$ are asymptotically normal. This follows from Theorem One and the fact that the Y's are sums of independent random variables.

Theorem One:

If $\{X_i\}$, $i\varepsilon I$, is a family of random variables, each independent and each distributed binomially with parameters p_i and N_i (for $0<p_i<1$ and $N_i>0$), where $\sum_{i\varepsilon I} p_i<\infty$, then for $W = \sum_{i\varepsilon I}(X_i-N_ip_i)/(\sum_{i\varepsilon I} N_ip_iq_i)^{\frac{1}{2}}$, the distribution of W asymptotically approaches that of a random variable which is distributed as normal (0,1) as $N_i\to\infty \forall i\varepsilon I$.

Proof: In order to prove the theorem, it is sufficient to show that the moment generating function, $G_W(m)$, approaches the limit $e^{\frac{1}{2}m^2}$ as $N_i\to\infty\forall i\varepsilon I$.

1. Let $B = (\sum_{i\varepsilon I} N_ip_iq_i)^{\frac{1}{2}}$

   a. The expansion of e^x is $\sum_{j=0}^{\infty} \frac{x^j}{j!}$ and for x small, is $1+x+\frac{x^2}{2} + o(x^2)$.

   b. For x small, the term $(1+x)^n$ can be represented by e^{nx}.

2. $G_W(m) = E(e^{mW}) = E\left(e^{m\sum_{i\varepsilon I}(X_i-N_ip_i)/B}\right) = \prod_{i\varepsilon I} E\left(e^{m(X_i-N_ip_i)/B}\right)$

$$= \prod_{i\varepsilon I}(p_ie^{m/B}+q_i)^{N_i} e^{-N_ip_im/B}$$

$$= \prod_{i\varepsilon I}\left(p_ie^{q_im/B} + q_ie^{-p_im/B}\right)^{N_i}.$$

3. From 1a, the expansion of $p_ie^{q_im/B}$ is $p_i[1 + q_im/B + \frac{1}{2}(q_im/B)^2 + o(1/B^2)]$ and the expansion of $q_ie^{-mp_i/B}$ is $q_i[1 - p_im/B + \frac{1}{2}(p_im/B)^2 + o(1/B^2)]$. Thus,

$$p_i e^{q_i m/B} + q_i e^{-p_i m/B} = (p_i+q_i) + (p_i q_i - p_i q_i)m/B$$
$$+ \tfrac{1}{2}p_i q_i (q_i+p_i)(m/B)^2 + o(1/B^2)$$
$$= 1 + \tfrac{1}{2}p_i q_i (m/B)^2 + o(1/B^2) .$$

4. From 1b, and using the fact that the term $\tfrac{1}{2}p_i q_i (m/B^2)$ is small for large values of B, and neglecting $o(1/B^2)$, we have

$$\left(p_i e^{q_i m/B} + q_i e^{-p_i m/B}\right)^{N_i} \sim e^{p_i q_i N_i (m/B)^2/2}, \text{ and hence}$$

$$G_w(m) \sim \prod_{i\varepsilon I} e^{\frac{1}{2}p_i q_i N_i (m/B)^2} = e^{\sum_{i\varepsilon I} \frac{1}{2}p_i q_i N_i (m/B)^2} = e^{\frac{1}{2}m^2}$$

in the limit as N_i tends to infinity for all i in I, since $(1/B)^2(\sum_{i\varepsilon I} p_i q_i N_i) = 1$. □

It is now possible to examine how best to estimate expected future values of Y and Y_i when past realizations of Y and Y_i are known. When the conditional expectation of a random variable cannot be found explicitly, Parzen (1960) suggests a best linear predictor which minimizes the mean square error of prediction using a linear function of previous realizations. When those random variables are normal, this function gives the exact conditional expectation of the random variable, given the value in the previous time periods. Theorem One justifies our use of the best linear predictor. In Theorem Two, we derive the general expression for this predictor, which we later specialize to our cohort model.

Theorem Two:

A best linear predictor is defined to be that function

$E^*(Y) = a + bX$, for Y and X random variables,

which minimizes $E[(Y-E^*[Y])^2]$, denoted $Var^*(Y)$. The expressions for $E^*(Y)$ and $VAR^*(Y)$ are

$E^*(Y) = E(Y) + [Cov(Y,X)/Var(X)][X-E(X)]$ and

$Var^*(Y) = Var(Y)\ (1-\rho^2(X,Y))$, where the term $\rho(X,Y)$ denotes the correlation coefficient of X,Y.

Proof:

1. $Var^*(Y) = E[(Y-(a+bX))^2]$.

2. $\frac{d}{da} Var^*(Y) = -2E[Y-(a+bX)] = -2E[Y] + 2(a+bE[X])$, which when set to zero, yields $a = E[Y] - bE[X]$.

3. $\frac{d}{db} Var^*(Y) = -2E[XY-X(a+bX)] = -2E[XY] + 2aE[X] + 2bE[X^2]$, which, when set to 0 and the substitution for a is made, yields $b = \frac{E[XY] - E[X]E[Y]}{E[X^2] - E^2[X]} = Cov[X,Y]/Var[X]$.

4. Substituting for a and b in the expression for $E^*(Y)$ and $Var^*(Y)$, $E^*(Y) = E[Y] - bE[X] + bX$

$= E[Y] + (Cov[X,Y]/Var[X])(X-E[X])$ and

$Var^*(Y) = E[(Y-E^*[Y])^2] = b^2E[(X-E[X]]^2) - 2bE[(X-E(X))(Y-E(Y))]$

$= Var[Y] + b^2Var[X] - 2bCov[X,Y]$

$= Var[Y] + [Cov(X,Y)/Var(X)]^2Var[X]$

$- 2(Cov[X,Y]/Var[X])Cov[X,Y]$

$= Var[Y][1-Cov^2[X,Y]/(Var[X]Var[Y])]$

$= Var[Y](1-\rho^2[X,Y])$. □

In a similar manner, $E^{**}(Y) = E(Y)+b_1(X-E(X))+b_2(Z-E(Z))$ minimizes $Var^{**}(Y) = E[(Y-E^{**}(Y))^2]$ when

$$b_1 = D[Cov(X,Y)/Var(X) - Cov(Y,Z)Cov(X,Z)/\{Var(X)Var(Z)\}],$$

$$b_2 = D[Cov(Z,Y)/Var(Z) - Cov(Y,X)Cov(X,Z)/\{Var(X)Var(Z)\}],$$

and

$$D = 1/(1-\rho^2[X,Z]).$$

The expression for $Var^{**}(Y)$ then becomes

$$Var^{**}(Y) = Var(Y) + b_1^2Var(X) + b_2^2Var(Z) + 2b_1b_2Cov(X,Z) - 2b_1Cov(Y,X) - 2b_2Cov(Y,Z).$$

We now have an expression for the best linear predictor for Y(t) when we know the past realizations, Y(t-1) and Y(t-2), regardless of the distribution of Y(t). For large cohort sizes at entry into the system, we have that this predictor <u>is</u> best (i.e., it is $E[Y(t)|Y(t-1),Y(t-2)]$), since the distribution of Y(t) will be very close to normal. Hence the functions E^* and E^{**} are actually conditional expectations given the past period error and the last two period errors respectively.

The only terms used in expressing E^* and E^{**} which have not been derived are the $Cov[Y(t),Y(t-1)]$ and $Cov[Y(t),Y(t-2)]$. Since independence exists between $X_i(t-y,y)$ and $X_j(t-y,y)$ for $i \neq j$, and independence exists between $X_i(t-s,s)$ and $X_i(t-u,u)$ for $u \neq s$, the covariance between all such X terms

is zero except for that between $X_i(t-s,u)$ and $X_i(t-s,w)$. Here, for $w>u$, we may interpret $X_i(t-s,u)$ to be the remnants of the initial size at entry, $X_i(t-s)$; and $X_i(t-s,w)$ to be the remnants, some w-u epochs later, of $X_i(t-s,u)$. Given that a member of $X_i(t-s)$ remains until a time u epochs later, the probability of his remaining until w epochs after entry into the system is $p_i(w)/p_i(u)$. (Assuming a person who leaves the system does not return, $0 \leq [p_i(w)/p_i(u)] < 1$ for $w>u$.) The conditional expectation of $X_i(t-s,w)$ given the realization of $X_i(t-s,u)$ is then $[p_i(w)/p_i(u)]X_i(t-s,u)$. Hence the covariance of $X_i(t-s,u)$ and $X_i(t-s,w)$ is derived by the following argument:

$$E[X_i(t-s,u),X_i(t-s,w)] = [p_i(w)/p_i(u)]E[X_i(t-s,u)^2];$$

$$\mathrm{Cov}[X_i(t-s,u),X_i(t-s,w)] = [p_i(w)/p_i(u)]E[X_i(t-s,u)^2] - E[X_i(t-s,u)]E[X_i(t-s,w)],$$

where

$$E[X_i(t-s,u)] = p_i(u)X_i(t-s) \text{ and}$$

$$E[X_i(t-s,w)] = p_i(w)X_i(t-s);$$

and

$$E[X_i(t-s,u)]E[X_i(t-s,w)] = p_i(w)p_i(u)[X_i(t-s)]^2$$

$$= [p_i(w)/p_i(u)][p_i(u)]^2[X_i(t-s)]^2$$

$$= [p_i(w)/p_i(u)]\{E[X_i(t-s)]\}^2.$$

Hence,

$$\begin{aligned}\mathrm{Cov}[X_i(t-s,u),X_i(t-s,w)] &= [p_i(w)/p_i(u)]E[X_i(t-s,u)^2] \\ &\quad -[p_i(w)/p_i(u)]E^2[X_i(t-s,u)] \\ &= [p_i(w)/p_i(u)]\mathrm{Var}[X_i(t-s,u)] \\ &= [p_i(w)/p_i(u)]p_i(u)q_i(u)X_i(t-s) \\ &= p_i(w)q_i(u)X_i(t-s).\end{aligned}$$

To express the covariance between $X_i(t-s,s)$ and the remaining number of the cohort with age s-1 (one period ago), $X_i(t-s,s-1)$, we substitute $w = s$ and $u = s - 1$ to obtain

$$\mathrm{Cov}[X_i(t-s,s),X_i(t-s,s-1)] = p_i(s)q_i(s-1)X_i(t-s), \text{ for } s \geq 1.$$

Similarly, by substituting $w = s$ and $u = s - 2$,

$$\mathrm{Cov}[X_i(t-s,s),X_i(t-s,s-2)] = p_i(s)q_i(s-2)X_i(t-s), \text{ for } s \geq 2.$$

To obtain $\mathrm{Cov}[Y_i(t),Y_i(t-1)]$, the covariance existing between all the remnants at time t from cohorts which entered grade i and the remnants from the same cohorts at time t-1, we sum on s from 1 to M to obtain,

$$\begin{aligned}\mathrm{Cov}[Y_i(t),Y_i(t-1)] &= \sum_{s=1}^{M} \mathrm{Cov}[X_i(t-s,s),X_i(t-s,s-1)] \\ &= \sum_{s=1}^{M} p_i(s)q_i(s-1)X_i(t-s).\end{aligned}$$

Similarly, (mutatis mutandi)

$$\mathrm{Cov}[Y_i(t),Y_i(t-2)] = \sum_{s=2}^{M} \mathrm{Cov}[X_i(t-s,s),X_i(t-s,s-2)]$$

$$= \sum_{s=2}^{M} p_i(s)q_i(s-2)X_i(t-s).$$

The independence between cohorts starting in different grades has already been established; thus, to obtain the covariance between all remnants at time t with all remnants at time t-1 and with all remnants at time t-2, we sum on i from 1 to K the covariances of the $Y_i(t)$ terms to obtain, respectively,

$$\mathrm{Cov}[Y(t),Y(t-1)] = \sum_{i=1}^{K} \mathrm{Cov}[Y_i(t),Y_i(t-1)]$$

$$= \sum_{i=1}^{K} \sum_{s=1}^{M} p_i(s)q_i(s-1)X_i(t-s)$$

and

$$\mathrm{Cov}[Y(t),Y(t-2)] = \sum_{i=1}^{K} \mathrm{Cov}[Y_i(t),Y_i(t-2)]$$

$$= \sum_{i=1}^{K} \sum_{s=2}^{M} p_i(s)q_i(s-2)X_i(t-s).$$

To summarize, the expressions for estimating the remaining persons in the system who started in grade i are

$$E^*(Y_i(t)) = E[Y_i(t)] + b_{i,t}\{Y_i(t-1) - E[Y_i(t-1)]\},$$

where

$$b_{i,t} = \mathrm{Cov}[Y_i(t),Y_i(t-1)]/\mathrm{Var}[Y_i(t-1)];$$

and

$$E^{**}(Y_i(t)) = E[Y_i(t)] + b^1_{i,t}\{Y_i(t-1) - E[Y_i(t-1)]\} + b^2_{i,t}\{Y_i(t-2) - E[Y_i(t-2)]\},$$

for

$$b^1_{i,t} = d(A_1 - A_2A_3) \quad \text{and} \quad b^2_{i,t} = d(A_3 - A_1A_2),$$

where

$$d = 1/\{1-\rho^2[Y_i(t-1),Y_i(t-2)]\},$$

$$A_1 = \mathrm{Cov}[Y_i(t),Y_i(t-1)]/\mathrm{Var}[Y_i(t-1)] = b_{i,t},$$

$$A_2 = \mathrm{Cov}[Y_i(t-1),Y_i(t-2)]/\mathrm{Var}[Y_i(t-2)] = b_{i,t-1}$$

and

$$A_3 = \mathrm{Cov}[Y_i(t),Y_i(t-2)]/\mathrm{Var}[Y_i(t-2)].$$

The expressions for estimating the number of total remaining persons are then

$$E^{*}(Y(t)) = E[Y(t)] + b_t\{Y(t-1) - E[Y(t-1)]\}$$

and

$$E^{**}(Y(t)) = E[Y(t)] + b^1_t\{Y(t-1) - E[Y(t-1)]\} + b^2_t\{Y(t-2) - E[Y(t-2)]\};$$

where

$$b_t = \mathrm{Cov}[Y(t),Y(t-1)]/\mathrm{Var}[Y(t-1)]$$

$$= \sum_{i=1}^{K} \mathrm{Cov}[Y_i(t),Y_i(t-1)] / \{\sum_{i=1}^{K} \mathrm{Var}[Y_i(t-1)]\},$$

$$b_t^1 = D\{b_t - b_{t-1}\mathrm{Cov}[Y(t),Y(t-2)]/\mathrm{Var}[Y(t-2)]\},$$

$$b_t^2 = D\{\mathrm{Cov}[Y(t),Y(t-2)]/\mathrm{Var}[Y(t-2)] - b_t b_{t-1}\},$$

and

$$D = 1/\{1 - \rho^2[Y(t-1),Y(t-2)]\}.$$

This model for predicting the total number in the system at time t is applied to the problem of predicting total student enrollment in Section V.

III. MEETING SPECIFIED GOALS

Consider an organization in which desired numbers are specified for personnel in each grade i at times t+1, t+2, ... in the future. It is the objective of this section to extend the model of Section II for predicting the future grade size within the organization.

Define $p_{ij}(t,u)$ to be that fraction of personnel in grade j at time u given they entered the system in grade i at time t, where t<u. If the grades are numbered in hierarchal order, i<j means a promotion, i>j means a demotion and i = j means that $p_{ii}(t,u)$ is the fraction which remain in grade i in the period t to u. If the rates of movement between grades are stable for the time period being considered, an assumption of stationarity or independence of time t can be made about the transferred fractions, and $p_{ij}(t,u)$ can be expressed as $p_{ij}(u-t)$; that is, the fraction transferred from grade i to j is a function only of the elapsed time u-t. Let $X_{ij}(t,u)$ denote the number in grade j at time u of that cohort which entered the organization at time t in grade i. We then have that $X_{ij}(t,u)$ is a binomially distributed random variable with parameters $X_i(t)$ and $p_{ij}(u-t)$.

By representing the number of persons in grade j at time t as $Y^j(t)$ in terms of the cohort portions remaining from the initial size at entry into grade i in all previous epochs, we sum on i and sum on s to obtain

$$Y^j(t) = \sum_{s=0}^{M} \sum_{i=1}^{K} X_{ij}(t-s,t).$$

It should be noted that $X_{ij}(t-s,t)$ is independent of $X_{ij}(t-r,t)$ for $s \neq r$, since the cohort entering grade i at time t-s acts independently of the cohort entering grade i at time t-r. Note also that $X_{kj}(t-s,t)$ is independent of $X_{ij}(t-s,t)$ for $i \neq k$, since these are the remnants presently in grade j at time t of those who entered the system at time t-s in different grades. For $s = 0$, $X_{ij}(t,t)$ is 0 when $i \neq j$ and $X_{ii}(t,t)$ is $X_i(t)$; that is, a cohort which just entered the system at time t in grade i will have no opportunity to diminish in size or to be transferred to another grade in the same time period. Of note is the fact that independence does not exist between $Y^j(t)$ and $Y^\ell(t)$ for $j \neq \ell$, since in each of these random variables, there may exist people from the same initial cohort. At this point, the distinction between $Y_i(t)$ and $Y^i(t)$ should be clear. In expressing $Y_i(t)$, we are counting the remnants of $X_i(t-s)$ in all grades. In expressing $Y^i(t)$, we are counting the number presently in grade i as remnants from all previous time periods of cohort entries into all grades.

We are now able to express the means and variances of $Y^j(t)$ as the sum of independent means and variances respectively of $X_{ij}(t-s,t)$;

$$E[Y^j(t)] = E[\sum_{i=1}^{K} \sum_{s=0}^{M} X_{ij}(t-s,t)] = \sum_{i=1}^{K} \sum_{s=0}^{M} p_{ij}(s)X_i(t-s)$$

and

$$\mathrm{Var}[Y^j(t)] = \sum_{i=1}^{K} \sum_{s=0}^{M} p_{ij}(s)q_{ij}(s)X_i(t-s) \text{ for } j=1,2,\ldots K;$$

where

$$p_{ii}(0) = 1, \text{ and } q_{ij}(s) = 1 - p_{ij}(s).$$

Denote by $G_i(t+1)$, $G_i(t+2)$,... the desired goals for the size of grade i at times t+1, t+2, ... respectively. Denote by $Z_i(t+1)$, $Z_i(t+2)$,... the numbers in grade i at times t+1, t+2,... which were in the system at time t, t+1,..., in one of the K grades of the organization. The number $X_i(t+1)$ is now a controlled variable; that is, the number to recruit into grade i at time t+1 in order to attain $G_i(t+1)$ is $X_i(t+1) = G_i(t+1) - Z_i(t+1)$.

We now assume that no demotions take place; $p_{ij}(s)$ is zero for $j<i$. At the end of each epoch, a member is promoted, remains in present grade or leaves the organization. With these restrictions placed on the system, we have that the expected values for $Z_i(t+1)$ and $Z_i(t+2)$ are

$$E[Z_i(t+1)] = \sum_{j=1}^{i} \sum_{s=0}^{M} X_j(t-s)\, p_{ji}(s)$$

and

$$E[Z_i(t+2)] = \sum_{j=1}^{i} \sum_{s=0}^{M} X_j(t-s+1)p_{ji}(s), \text{ for } i = 1,2,\ldots K.$$

The variance terms are

$$\mathrm{Var}[Z_i(t+1)] = \sum_{j=1}^{i} \sum_{s=0}^{M} X_j(t-s)p_{ji}(s)q_{ji}(s)$$

and

$$\mathrm{Var}[Z_i(t+2)] = \sum_{j=1}^{i} \sum_{s=0}^{M} X_j(t-s+1)p_{ji}(s)q_{ji}(s).$$

With the same definition as in Section II for $p_i(s)$ (that is, the fraction of $X_i(t-s)$ remaining in the system at time t), we have a restriction that $\sum_{j=1}^{K} p_{ij}(s) = p_i(s)$ for all values of s (i.e., s = 0,1,...M).

Suppose our problem is to avoid overmanning or undermanning grade i (i=1,2,...K) in the periods t+1, t+2, when the recruiting for these periods must be planned ahead. Assume that there are penalties defined when the number of personnel in each grade is above or below the goals; i.e., $C_i^+(t)$ when the number is over the desired values and $C_i^-(t)$ when below. The objective function to be minimized is

$$E_Z\left[\sum_{x=1}^{2}\sum_{i=1}^{K}\{C_i^+(t+x)\mathrm{Max}[0,G_i(t+x)-X_i(t+x)-Z_i(t+x)] - C_i^-(t+x)\mathrm{Min}[0,G_i(t+x)-X_i(t+x)-Z_i(t+x)]\}\right],$$

where $E_Z(\cdot)$ denotes the expected value of $(\cdot)$ over the joint density of $Z = (Z_1(t+1),Z_2(t+1),\ldots Z_K(t+1),Z_1(t+2),\ldots Z_K(t+2))$, a vector of random variables with a multivariate normal distribution. The optimal values of $X_i(t+1)$ and $X_i(t+2)$

would minimize this objective function. When $C_i^+ = C_i^- = C_i$, the expression reduces to $\sum_{x=1}^{2} \sum_{i=1}^{K} C_i(t+x) \left| \{G_i(t+x) - X_i(t+x) - E[Z_i(t+x)]\} \right|$, where the expected values of $Z_i(t+1)$ and $Z_i(t+2)$ are given above.

We now assume that the realizations, $Y_i(t)$ i=1,2,...K are known and proceed to determine the best predictors for $Z_i(t+1)$ and $Z_i(t+2)$. Assuming the normality property for Z_i and Y_i, this best predictor will be the expectation of Z_i conditioned on the values of Y_j, j=1,2,...i, for the present time; defining $E'[Z_i(t+1)]$ as the best predictor of $Z_i(t+1)$ and $E''[Z_i(t+2)]$ as the best predictor of $Z_i(t+2)$ when the values of $Y_j(t)$ are known, j=1,2,...K, we have

$$E'[Z_i(t+1)] = E[Z_i(t+1) \mid Y_j(t),\ j=1,2,...i]$$

and

$$E''[Z_i(t+2)] = E[Z_i(t+2) \mid Y_j(t),\ j=1,2,...i]$$

where the realizations of $Y_j(t)$ for j>i have no effect on E' and E'' and are therefore not considered in the expressions. (This follows from the fact that $p_{ji}(s)$ is zero for i<j.)

With the independence of $Y_i(t)$ and $Y_j(t)$ for $i \neq j$, we have that $Cov[Y_i(t),Y_j(t)] = 0$ for $i \neq j$. Since $X_i(t+1)$ is a control variable, we also have that $Var[Y^j(t+x)] = Var[Z_j(t+x)]$ and $Cov[Y^j(t+x),Y_i(t)] = Cov[Z_j(t+x),Y_i(t)]$ for x=1,2 and i=1,2,...K. Extending Theorem Two of Section II to E' and E'' we thus have that

$$E'(Z_i(t+1)) = E[Z_i(t+1)] + \sum_{j=1}^{i} d_t^{ji}\{Y_j(t) - E[Y_j(t)]\},$$

and

$$E''(Z_i(t+2)) = E[Z_i(t+2)] + \sum_{j=1}^{i} g_t^{ji}\{Y_j(t) - E[Y_j(t)]\},$$

where

$$d_t^{ji} = Cov[Y^i(t+1),Y_j(t)]/Var[Y_j(t)]$$

and

$$g_t^{ji} = Cov[Y^i(t+2),Y_j(t)]/Var[Y_j(t)]$$

for $i = 1,2,\ldots K$. The terms $E[Z_i(t+x)]$, $x=1,2$, and $Var[Y_j(t)]$ have been derived previously, leaving the terms $Cov(Y^i(t+x),Y_j(t))$, $x=1,2$, to be derived:

Given that a member who entered the organization at time t-s in grade j was in the organization at time t, (i.e. a member of $X_j(t-s,s)$), the probability that this member will be in grade i at time t+x, is $p_{ji}(s+x)/p_j(s)$. The conditional expectation $E[X_{ji}(t-s,s+x)|X_j(t-s,s)]$ (that is the expectation of the number in grade i at time t+x of those who started in grade j at time t-s, conditioned on the number remaining in the organization at time t who started at time t-s in grade j), is $p_{ji}(s+x)X_j(t-s,s)/p_j(s)$. Using a similar argument as that in deriving $Cov[Y(t),Y(t-1)]$ in Section II, we have that

$$E[X_{ji}(t-s,s+x),X_j(t-s,s)] = \frac{p_{ji}(s+x)}{p_j(s)} E[(X_j(t-s,s))^2]$$

hence

$$Cov[X_{ji}(t-s,s+x),X_j(t-s,s)] = \frac{p_{ji}(s+x)}{p_j(s)} Var[(X_j(t-s,s)]$$

$$= p_{ji}(s+x)q_j(s)\ X_j(t-s).$$

Since independence exists between $X_j(t-s,s)$ and $X_j(t-r,r)$ and between $X_{ji}(t-s,s+x)$ and $X_{ji}(t-r,r+x)$ for $s \neq r$, we have

$$Cov[Y^i(t+x),Y_j(t)] = \sum_{s=0}^{M-x} Cov[X_{ji}(t-s,s+x),X_j(t-s)]$$

$$= \sum_{s=0}^{M-x} p_{ji}(s+x)q_j(s)X_j(t-s),$$

$$\text{for } x = 1,2;\ \ j=1,2,\ldots i.$$

The use of E' and E'' in place of $E(Z_i(t+1))$ and $E(Z_i(t+2))$ respectively (using the present values of $Y_i(t)$), is relevant in the objective function when $C_i(t+x) = C_i^+(t+x) = C_i^-(t+x)$, $x=1,2$. The variances (using the fact that $Cov(Y_i(t),Y_j(t))$ is zero for $i \neq j$), are

$$Var'[Z_i(t+1)] = Var[Z_i(t+1)] + \sum_{j=1}^{i} \{(d_t^{ji})^2 Var[Y_j(t)]$$

$$- 2d_t^{ji} Cov[Y^i(t+1),Y_j(t)]\}$$

and

$$\mathrm{Var}''[Z_i(t+2)] = \mathrm{Var}[Z_i(t+2)] + \sum_{j=1}^{i} \{(g_t^{ji})^2 \mathrm{Var}[Y_j(t)] - 2g_t^{ji}\mathrm{Cov}[Y^i(t+2), Y_j(t)]\}.$$

An application of such a model might be that of a university system with the grades defined for curriculum and level (upper and lower, for instance) in which the goals $G_i(t+1)$ and $G_i(t+2)$ were specified to fully utilize the facilities without inflicting a lack of classrooms by over-enrolling. The costs, $C_i(t+1)$ and $C_i(t+2)$ might be based on the losses encurred financially by overstaffing for a below-desired-level of class size and the losses encurred by the added administrative burdens of rejecting enrolled students when overages occur.

IV. SENSITIVITY TO ERRORS IN PROBABILITY ESTIMATION

The fraction of remaining individuals who start in a grade and remain for s epochs, s=1,2,...M is used as an estimate of the probability of this event. In the paper by McAfee (1970), a statistical test is made on a sample of size three, to test the hypothesis that $p_i(s)$ for each cohort is from the same population. This is to say that for moderate sample sizes in estimating $p_i(s)$, we have a sample mean to use, which for large sample sizes, approaches the true value of $p_i(s)$. At best, we use the estimate and for this reason examine the sensitivity of the model described in Section II to error which may exist between our estimate and the true value of $p_i(s)$; viz., the sensitivity of the expected values, E[Y(t)], E*[Y(t)] and E**[Y(t)] to errors in $\hat{p}_i(s)$. We shall consider two cases: one, in which an error $\Delta p_i(s)$ exists between our estimate and the true value of $p_i(s)$ for some s and i; the second, in which an error exists for all s and i.

Taking the partials of E[Y(t)], E*[Y(t)] and E**[Y(t)] with respect to $p_i(s)$ yields the following expressions (where $\frac{d(\cdot)}{dp_i(s)}$ denotes the partial differential of $(\cdot)$ with respect to $p_i(s)$):

$$dE[Y(t)]/dp_i(s) = d/dp_i(s)\ [X_i(t-s)p_i(s)] = X_i(t-s);$$

$$dE^*[Y(t)]/dp_i(s) = d/dp_i(s)\, E[Y(t)]$$
$$+b_t d/dp_i(s)\, \{Y(t-1) - E[Y(t-1]\}$$
$$+ \{Y(t-1) - E[Y(t-1)]\}d/dp_i(s)(b_t)$$
$$= X_i(t-s) - b_t X_i(t-s-1) + R_Y^*;$$

and

$$dE^{**}[Y(t)]/dp_i(s) = d/dp_i(s)E[Y(t)]$$
$$+ b_t^1 d/dp_i(s)\{Y(t-1) - E[Y(t-1)]\}$$
$$+ b_t^2 d/dp_i(s)\{Y(t-2) - E[Y(t-2)]\}$$
$$+\{Y(t-1) - E[Y(t-1)]\}d/dp_i(s)(b_t^1)$$
$$+\{Y(t-2) - E[Y(t-2)]\}d/dp_i(s)(b_t^2)$$
$$= X_i(t-s) - b_t^1 X_i(t-s-1) - b_t^2 X_i(t-s-2)+R_Y^{**};$$

where R_Y^* and R_Y^{**} represent sums of remainder terms negligible in comparison to other terms in the expressions respectively for $dE^*/dp_i(s)$ and $dE^{**}/dp_i(s)$. (See the appendix for the exact expressions represented by R.)

Neglecting R for small errors in the estimate for $p_i(s)$, the changes in predicted values are then

$$\Delta E[Y(t)] = X_i(t-s)\ \Delta p_i(s),$$

$$\Delta E^*[Y(t)] = \{X_i(t-s) - b_t X_i(t-s-1)\}\ \Delta p_i(s),$$

and

$$\Delta E^{**}(Y(t)) = \{X_i(t-s) - b_t^1 X_i(t-s-1) - b_t^2 X_i(t-s-2)\}\Delta p_i(s).$$

It is observed in the expression for $E^*(Y(t))$, that as b_t approaches one, the error in the predicted value is $\Delta p_i(s)\Delta X_i$, where $\Delta X_i = X_i(t-s) - X_i(t-s-1)$. Since the covariance of $Y(t)$ and $Y(t-1)$ is always positive, b_t takes on a value between zero and one, and b_t acts as a dampening coefficient for errors in the predicted value of $Y(t)$ using E^*.

Similarly for the expression of $\Delta E^{**}(Y(t))$, as $(b_t^1 + b_t^2) \to 1$, the error in the predicted value is $\Delta p_i(s)(b_t^1 \Delta X_i^1 + b_t^2 \Delta X_i^2)$, where the terms $\Delta X_i^1 = X_i(t-s) - X_i(t-s-1)$ and $\Delta X_i^2 = X_i(t-s) - X_i(t-s-2)$. (Viz., the differences between the cohort entering grade i at time t-s and those which enter the periods before.) The dampening effect of the error in $E^{**}[Y(t)]$ is evident with the b_t^i coefficients $(i=1,2)$. Thus, a smaller error in the predicted value of $Y(t)$ results when E^* and E^{**} are used rather than E.

Let us assume now that $p_i(s) = (p_i)^s$ for all i; i.e., we have a geometric distribution of remaining members of each grade. If $X_i(t)$ were a constant for all t (that is, $X_i(t) = X_i$), then

$$E[Y(t)] = \sum_{i=1}^{K} E[Y_i(t)] = \sum_{i=1}^{K} \left[\sum_{s=0}^{\infty} X_i p_i^s \right] = \sum_{i=1}^{K} X_i/(1-p_i);$$

$$E^*(Y(t)) = \sum_{i=1}^{K} X_i/(1-p_i) + b_t\{Y(t-1) - X_i/(1-p_i)\},$$

for

$$b_t = b = \left[\sum_{i=1}^{K} p_i^2 X_i/(1-p_i^2) \right] / \left[\sum_{i=1}^{K} p_i X_i/(1-p_i^2) \right];$$

and

$$E^{**}(Y(t)) = \sum_{i=1}^{K} X_i/(1-p_i) + b_t^1[Y(t-1) - X_i/(1-p_i)]$$

$$+ b_t^2[Y(t-2) - X_i/(1-p_i)],$$

for

$$b_t^1 = [b/(1-b)][1 - \sum_{i=1}^{K} p_i{}^3X_i / \sum_{i=1}^{K} p_iX_i)] = b^1$$

and

$$b_t^2 = [1/(1-b)]\left\{\left[\sum_{i=1}^{K} p_i{}^3X_i/(\sum_{i=1}^{K} p_iX_i\right] - (b)^2\right\} = b^2.$$

The resultant terms for the changes in prediction are then

$$\Delta E(Y(t)) = \sum_{i=1}^{K} \Delta p_iX_i/(1-p_i{}^2),$$

$$\Delta E^*(Y(t)) = \sum_{i=1}^{K} \Delta p_iX_i(1-b)/1-p_i{}^2) = (1-b)\Delta E(Y(t)),$$

and

$$\Delta E^{**}(Y(t)) = \sum_{i=1}^{K} \Delta p_iX_i(1-b^1-b^2)/(1-p_i{}^2) = (1-b^1-b^2)\Delta E(Y(t)).$$

The dampening factors for $\Delta E^*(Y(t))$ and $\Delta E^{**}(Y(t))$ are immediately evident.

Using E^* or E^{**} in place of E, we can reduce the prediction error caused by errors in $\hat{p}_i(s)$.

V. APPLICATION TO UNIVERSITY ENROLLMENT

The problem of predicting total student attendance at the University of California, Berkeley, is approached using the model for the best predictors, E*(Y(t)) and E**(Y(t)). Data was obtained from Berkeley for new student enrollments and is tabulated in Table I.

Total Student predictions are required for only the fall quarter of each year. To estimate the $p_i^j(s)$ (the probability of a student remaining to the s^{th} subsequent fall period following entry into grade i, quarter j), data was collected on the numbers of students from the cohorts entering in the Fall of 1966 and the Winter, Spring and Summer of 1967 which were still in attendance each succeeding fall term. The most recent data available was from 1969, which included at most 3 years for any cohort. To estimate $p_i^j(s)$ for the years in attendance 4 - 6, we assumed students' attendance behavior over time is essentially stationary and used past cohort data analysis found in Suslow et al (1968). Our estimates of $p_i^j(s)$ are given in Table II. It should be pointed out that during the years 1961 through 1966, the University followed a semester system. Starting in 1967, the University switched to a quarter system.

The parameters calculated with the data in Tables I and II are shown in Table III. Our estimates of total student

TABLE I

NEW ENROLLMENTS AT THE UNIVERSITY OF CALIFORNIA, BERKELEY

Year	FALL				SPRING			
	Freshman	Sophomore	Junior	Senior	Freshman	Sophomore	Junior	Senior
1961	3473	751	1160	146	312	171	263	29
1962	3528	678	1416	186	324	204	303	30
1963	3632	732	1568	196	328	187	324	42
1964	3443	609	1443	303	346	209	173	45
1965	2590	396	1035	126	256	180	452	49
1966	3048	720	1397	200	291	210	476	66
1967	3293	838	1648	170	255	161	300	39
1968	2234	534	1354	122	214	204	361	37
1969	1872	257	809	47	259	194	501	60

Year	WINTER				SUMMER			
	Freshman	Sophomore	Junior	Senior	Freshman	Sophomore	Junior	Senior
1961								
1966	0	0	0	0	0	0	0	0
1967	95	124	176	46	175	104	259	72
1968	111	143	251	48	409	246	517	205
1969	193	254	507	84	915	241	640	176

TABLE II

FRACTIONS OF STUDENTS ATTENDING EACH SUCCESSIVE FALL AFTER ENROLLMENT

Year	FRESHMAN ENROLLING IN				SOPHOMORE ENROLLING IN			
	Fall	Winter	Spring	Summer	Fall	Winter	Spring	Summer
0	1	.83	.85	.86	1	.79	.82	.71
1	.78	.62	.695	.61	.72	.63	.63	.59
2	.665	.505	.62	.53	.51	.205	.33	.33
3	.605	.245	.405	.45	.143	.025	.045	.045
4	.3	.062	.13	.2	.0175	0	0	0
5	.075	.016	.032	.05				
6	.019	0	0	0				

Year	JUNIOR ENROLLING IN				SENIOR ENROLLING IN			
	Fall	Winter	Spring	Summer	Fall	Winter	Spring	Summer
0	1	.79	.82	.67	1	.65	.74	.565
1	.72	.19	.465	.45	.3	.087	.23	.225
2	.18	.04	.087	.135	.11	.043	.077	.069
3	.045	0	0	0	.03	0	0	0

TABLE III

VALUES USED IN PREDICTION

CASE A
(Summer enrollees treated as Fall enrollees)

	Year			
	1966	1967	1968	1969
$E(Y(t))$	16919.5	18339.5	18365.5	18133.7
$Var(Y(t))^{\frac{1}{2}}$	67.66	67.86	71.02	70.83
$Cov(Y(t),Y(t-1))$		2261.1	2186.0	2349.6
$Cov(Y(t),Y(t-2))$			984.5	970.6
$\rho(Y(t),Y(t-1)$		.49245	.45358	.46707
b_t		.49400	.47471	.46586
b_t^1			.48641	.46055
b_t^2			-.02561	-.01305
$Y(t)-E(Y(t))$	-172.5	-2.518	-374.49	-17.742
$Var^*(Y(t))^{\frac{1}{2}}$		59.06	63.29	62.63
$Var^{**}(Y(t))^{\frac{1}{2}}$			63.49	62.63

CASE B
(Summer enrollees treated separately)

	Year			
	1966	1967	1968	1969
$E(Y(t))$	16919.5	18186.9	17848.4	17329.6
$Var(Y(t))^{\frac{1}{2}}$	67.66	68.73	73.04	73.60
$Cov(Y(t),Y(t-1))$		2177.6	2344.6	2398.7
$Cov(Y(t),Y(t-2))$			971.9	1077.6
$\rho(Y(t),Y(t-1))$		.46828	.46704	.44559
b_t		.47577	.49637	.44955
b_t^1			.51040	.44672
b_t^2			.03070	.00637
$Y(t)-E(Y(t))$	-172.5	150.12	142.61	786.42
$Var^*(Y(t))^{\frac{1}{2}}$		60.75	64.59	65.89
$Var^{**}(Y(t))^{\frac{1}{2}}$			64.74	65.78

enrollment in the falls of 1967 through 1969 are given in Table IV with actual enrollment figures for comparison. Due to a quota limit for fall enrollees and the start of year round operations in the Summer of 1968, it was felt that summer enrollees in 1968 and 1969 might not behave as summer enrollees in 1967, but in fact be early fall applicants who enrolled in the summer rather than risk unsuccessful enrollment in fall. This fact plus the relatively small sample size from 1967 for estimating the probabilities for summer enrollees to remain in the system, leads to two cases for estimation: Case A, in which summer enrollees are treated as fall enrollees; and Case B, in which summer enrollees are considered separately from the fall enrollees. The two cases are tabulated in Tables III and IV.

TABLE IV

PREDICTED TOTAL FALL ENROLLMENT
(With plus or minus two standard deviations)

CASE A
(Summer Enrollees treated as Fall Enrollees)

	Year		
	1967	1968	1969
Estimate using E	18339±135	18365±142	18133±141
Estimate using E*	18256±118	18364±126	17959±125
Estimate using E**		18368±126	17961±125

CASE B
(Summer Enrollees treated separately)

	Year		
	1967	1968	1969
Estimate using E	18186±137	17848±146	17329±147
Estimate using E*	18104±122	17992±129	17393±131
Estimate using E**		17919±129	17394±131
Actual Total Enrollment	18337	17991	18116

APPENDIX A

REMAINDER TERMS TO PREDICTION ERROR

In Section IV, expressions for $dE^*/dp_i(s)$ and $dE^{**}/dp_i(s)$ are given in which small remainder terms are represented by R_Y^* and R_Y^{**} respectively. The term R_Y^* represents $\{Y(t-1) - E[Y(t-1)]\}db_t/dp_i(s)$ where

$$db_t/dp_t(s) = \frac{\{X_i(t-s-1)q_i(s-1) - X_i(t-s)p_i(s+1) - b_t X_i(t-s-1)[1-2p_i(s)]\}}{Var[Y(t-1)]}.$$

The term R_Y^{**} represents the sum,

$$\{Y(t-1)-E[Y(t-1)]\}db_t^1/dp_i(s) + \{(Y(t-2)-E[Y(t-2)]\}db_t^2/dp_i(s).$$

$$db_t^1/dp_i(s) = \frac{\Delta^2}{Var[Y(t-1)]}\{b_t Var[Y(t-1)] - b_{t-1}\frac{Cov[Y(t),Y(t-2)]}{Var[Y(t-2)]}\}(D_1-D_2)$$

$$+ \frac{\Delta}{Var[Y(t-1)]}(A_1-A_2-A_3-A_4+A_5+A_6),$$

$$db_t^2/dp_i(s) = \frac{\Delta^2}{Var[Y(t-2)]}\{Cov[Y(t),Y(t-2)] - b_t b_{t-1} Var[Y(t-1)]\}(D_1-D_2)$$

$$+ \frac{\Delta}{Var[Y(t-2)]}(C_1-C_2-C_3-C_4),$$

where

$$A_1 = \frac{1}{Var[Y(t-1)]}[X_i(t-s-1)q_i(s-1) - X_i(t-s)p_i(s+1)]$$

$$A_2 = b_t X_i(t-s-1)[1-2p_i(s)]$$

$$A_3 = b_{t-1}[X_i(t-s-2)q_i(s-2) - X_i(t-s)p_i(s+2)]$$

$$A_4 = \frac{Cov[Y(t),Y(t-2)]}{Var[Y(t-2)]}\{X_i(t-s-2)q_i(s-1) - X_i(t-s-1)p_i(s+1)\}$$

$$A_5 = \frac{Cov[Y(t),Y(t-2)]}{Var[Y(t-2)]}\frac{Var[Y(t)]}{Var[Y(t-2)]}b_{t-1}X_i(t-s-2)[1-2p_i(s)]$$

$$A_6 = \frac{Cov[Y(t),Y(t-2)]}{Var[Y(t-2)]}b_{t-1}X_i(t-s)[1-2p_i(s)]$$

$$D_1 = X_i(t-s-2)\{2q_i(s-1) + b_{t-1}[2q_i(s-1)]\}$$

$$D_2 = X_i(t-s-1)\left[2p_i(s+1) + b_{t-1}\{2q_i(s-1)\}\frac{Var(Y(t-1))}{Var(Y(t-2))}\right]$$

$$C_1 = \frac{1}{Var[Y(t-2)]}\{X_i(t-s-2)q_i(s-2) - X_i(t-s)p_i(s+2)\}$$

$$C_2 = \frac{Cov[Y(t),Y(t-2)]}{Var[Y(t-2)]}X_i(t-s-2)[1-2p_i(s)]$$

$$C_3 = b_t[X_i(t-s-2)\{q_i(s-1) - b_{t-1} + 2p_i(s)\} - X_i(t-s-1)p_i(s+1)]$$

$$C_4 = b_{t-1}[X_i(t-s-1)\{q_i(s-1) - b_t + 2p_i(s)\} - X_i(t-s)p_i(s+1)]$$

and

$$\Delta = 1/(1-\rho^2[Y(t-1),Y(t-2)]).$$

BIBLIOGRAPHY

1. Bartholomew, D. J., Stochastic Models for Social Processes, John Wiley and Sons, London, 1967.

2. McAfee, K. M., A Cohort Model for Predicting Retention of Regular Marine Corps Officers, Master of Science in Operations Research Thesis, Naval Postgraduate School, Monterey, March 1970.

3. Marshall, K. T., and Oliver, R. M., A Constant Work Model for Student Attendance and Enrollment, Research Report No. 69-1, Office of the Vice President-Planning and Analysis, University of California, February 1969.

4. Parzen, E., Modern Probability Theory and Its Applications, John Wiley and Sons, New York, 1960.

5. Suslow, S., Langlois, E., Sumariwall, R., and Walther, C., Student Performance and Attrition at the University of California, Berkeley, Office of Institutional Research, University of California, Berkeley, January 1968.

6. Thonstad, T., Education and Manpower, University of Toronto Press, Toronto, 1968.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Documentation Center Cameron Station Alexandria, Virginia 22314	20
2.	Library, Code 0212 Naval Postgraduate School Monterey, California 93940	2
3.	Asst. Professor K. T. Marshall, Code 55Mt Department of Operations Analysis Naval Postgraduate School Monterey, California 93940	10
4.	Lieutenant Commander Raymond D. Hager 9226 Cayuga Drive Niagara Falls, New York 14340	1
5.	Bureau of Personnel, Code Pers B1401 Department of the Navy Washington, D.C. 22134	1
6.	Personnel Research Laboratory Department of the Navy Washington, D.C. 22134	1
7.	Chief of Naval Operations, Code OP96 Department of the Navy Washington, D.C. 22134	1
8.	Dr. R. M. Oliver Administrative Studies Program Office of Institutional Research Room 210 - T8 University of California Berkeley, California 94720	1
9.	Department of Operations Analysis Naval Postgraduate School Monterey, California 93940	1

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Naval Postgraduate School
Monterey, California 93940

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE

ANALYSIS OF A COHORT PREDICTION MODEL WITH APPLICATIONS TO STUDENT ENROLLMENT FORECASTING

4. DESCRIPTIVE NOTES (Type of report and, inclusive dates)
Master's Thesis; April 1970

5. AUTHOR(S) (First name, middle initial, last name)
Raymond D. Hager, Jr., Lieutenant Commander, United States Navy

6. REPORT DATE
April 1970

7a. TOTAL NO. OF PAGES
43

7b. NO. OF REFS
6

8a. CONTRACT OR GRANT NO.

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Naval Postgraduate School
Monterey, California 93940

13. ABSTRACT

A model is presented for the prediction of future organization size based on the numbers of recruits entering the organization in the past. This model utilizes the correlation between populations of successive time periods in order to better estimate the future remaining personnel in the system. The number of personnel leaving from each recruit cohort is assumed to follow the same probability distribution, which is a function of the age of the cohort in the organization and the grade in which the cohort started. For large cohort sizes the total personnel in the system is approximately normally distributed. This result justifies the use of a best linear prediction method which takes into account past errors of estimating the continuing population from one period to the next. Sensitivity of predictions to errors in probability estimates is discussed. The model is applied to predicting university student enrollment. Comparison of predicted and actual student enrollment is included.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
COHORT PREDICTION MODEL						
PERSONNEL RETENTION PREDICTION						
BEST LINEAR PREDICTOR						
UNIVERSITY STUDENT ENROLLMENT PREDICTION						